Simple Regression & Correlation

Numerical Variables

The difference between an economist and a statistician: people believe what economists say about the future, but not what statisticians say about the past - Henry Bottomley

The best thing about being a statistician is that you get to play in everyone's backyard.

- J.W. Tukey

If I had only one day left to live, I would live it in my statistics class – it would seem so much longer.

- Quote in a university student calendar

Learning Objectives

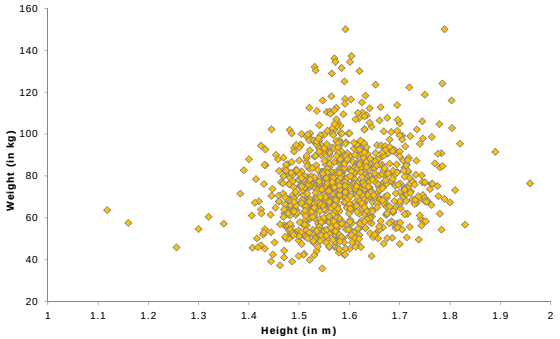
- To use scatter diagram to visualize relationship between two variables
- Use least squares method to develop simple linear regression models
- To learn measures of strength of association between two variables
  - Coefficient of determination
  - Correlation coefficient
- To learn uses and misuses of regression

Correlation Analysis

- A statistical tool used to describe how much one variable is linearly related to another - a way of measuring the extent to which two variables are related
- It measures the pattern of responses across variables

Scatter Diagram (Scatter Plot)

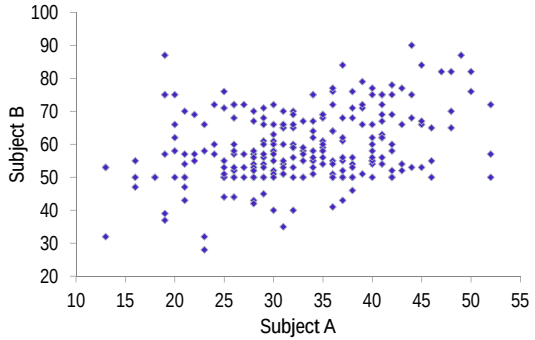
showing a weak positive relationship



Corrie Uys
UysC@cput.ac.za

Scatter Diagram (Scatter Plot)

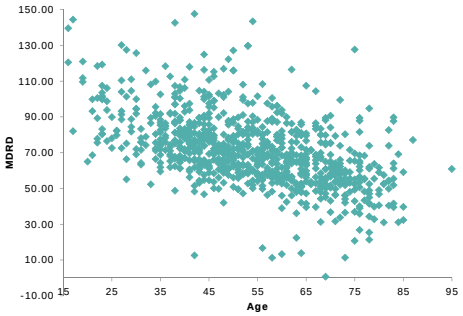
showing a stronger relationship



Corrie Uys
UysC@cput.ac.za

Scatter Diagram (Scatter Plot)

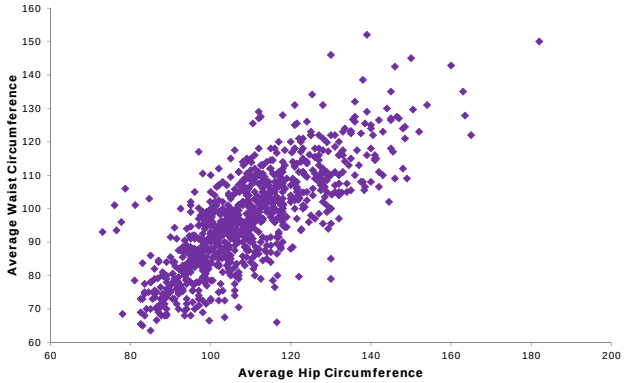
showing a stronger negative relationship



Corrie Uys
UysC@cput.ac.za

Scatter Diagram (Scatter Plot)

showing a strong relationship

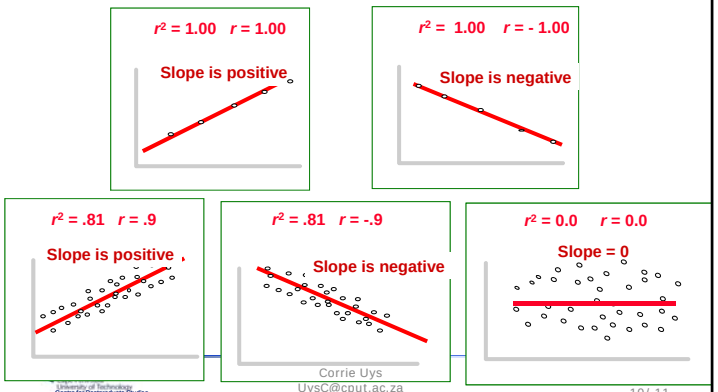


Corrie Uys
UysC@cput.ac.za

Coefficient of Correlation (r)

- Measures the strength of a linear relationship between two variables
- Sign of the slope (b_1) is the sign of r
- $-1 \leq r \leq 1$

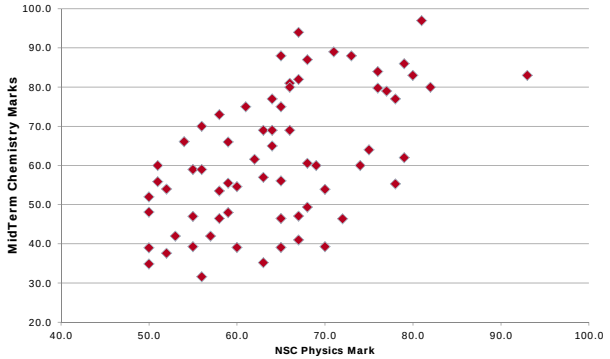
Various *r* Values



Example

A lecturer wanted to know if there was a relationship between the midterm first year chemistry marks and the nation senior certificate physics marks of first year chemical engineering students

Scatter plot



Practise

- Open Chem.Eng Regression Example.sav

Coefficient of Correlation (r)

Subject Example

$$r = +0.563$$

There seems to be a positive correlation between NSC Physics and Midterm Chemistry.

Is this significant, though?

Correlation Hypothesis Test

H_0: There is no relationship between NSC Physics and MidTerm Chemistry

H_1: There is a relationship between NSC Physics and MidTerm Chemistry.

	NSC_PhysicsPerc
Pearson Correlation	0.563
p-value. (2-tailed)	0.000
N	67

Conclusion: Since p-value < 0.05, we reject H_0 and conclude that there is a significant relationship between NSC Physics and MidTerm Chemistry

Various Correlation Coefficients

Parametric data:

- Pearson's r - when both variable are interval or ratio

Non-parametric data

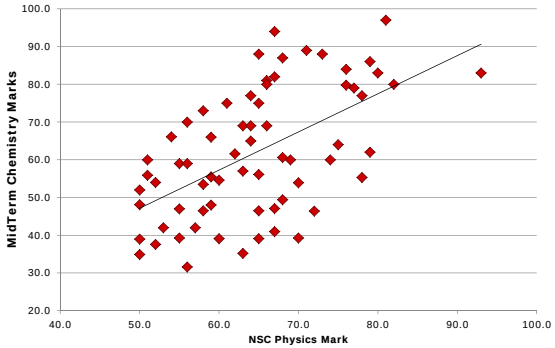
- Spearman's rank correlation or Spearman's ρ (rho)
  - Smaller datasets
  - Data can be ordinal or can be ranked
  - Based on differences in ranks between each of the pairs of ranks
- Kendall's Tau
  - Better than Spearman's ρ for small samples

Regression *Models*

1. Answer to 'What is the relationship between the variables?'
2. Equation used
   - One Numerical dependent variable
     - What Is to Be Predicted
   - One or More numerical or categorical independent (explanatory) variables

Definition of Terms

- Variables
  - Dependent (*Y*)
  - Independent (*X*)
- Scatter diagram - graphical representation of the (*X*,*Y*) pairs of observations
- Regression analysis - provides a "best fit" mathematical equation (i.e., relationship is expressed in equation form)



Definitions ... continued

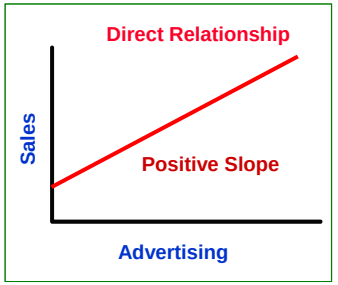
- Simple Linear Regression
  - Simple - only one independent or predictor variable (*X*)
  - Linear - the mathematical relation between *X* and *Y* is in the form:

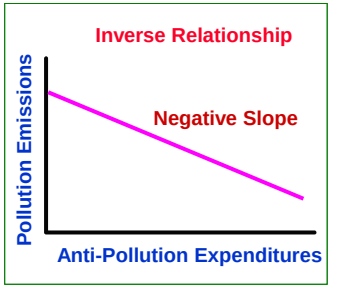
$$Y_i = b_0 + b_1 X_i$$

Types of Relationships

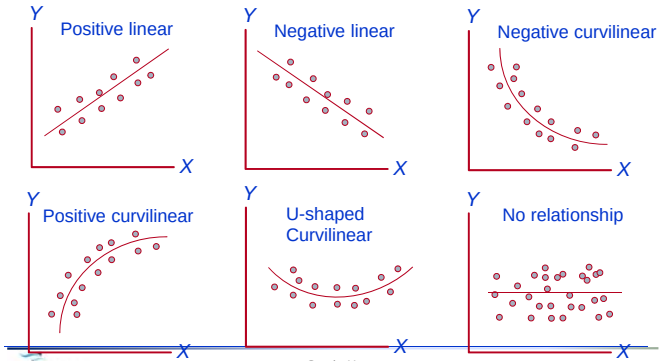
- Direct vs. Inverse
  - Direct - *X* and *Y* increase together
  - Inverse - *X* and *Y* have opposite directions
- Linear vs. Curvilinear
  - Linear - Straight line best describes the relationship between *X* and *Y*
  - Curvilinear - Curved line best describes the relationship between *X* and *Y*

Direct vs. Inverse Relationship

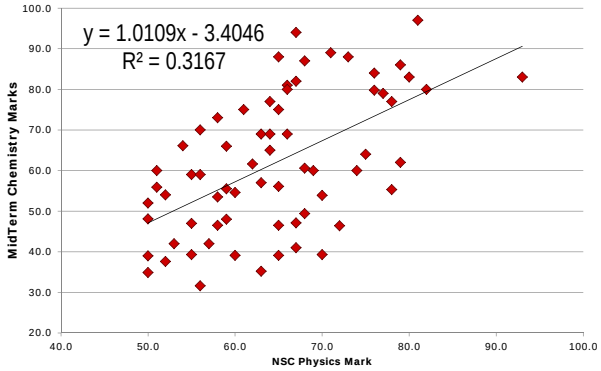




Possible Relationships Between X and Y in Scatter Diagrams



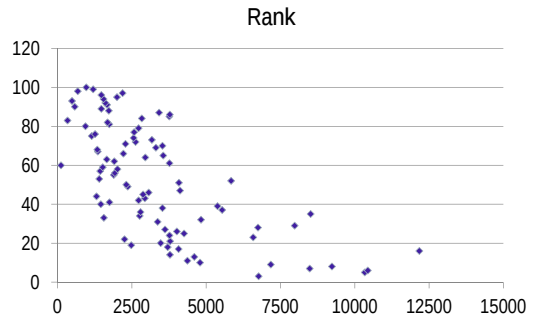
Adding the regression line



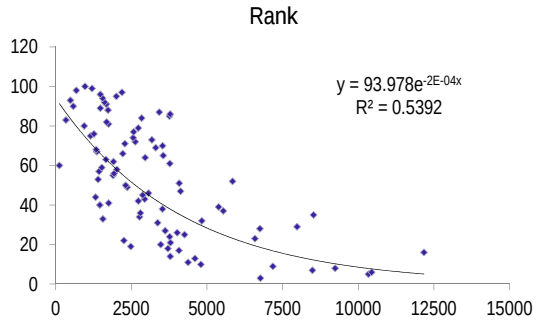
Practise

- Open Chem.Eng Regression Example.sav

Non-linear relationship



Non-linear relationship



R-squared = Coefficient of Determination

- Measures the proportion of variation explained by the independent variable in regression model
- For chemistrydata $R^2 = 0.563^2 = 0.3167$
  Thus 31.67% of the variation in the marks of midterm chemistry is directly attributed to the variation in the marks of NSC Physics
- Is subject 2 dependent on subject 1?

Hypothesis Test – F Test

H_0: There is no relationship between NSC Physics and MidTerm Chemistry

H_1: There is a relationship between NSC Physics and MidTerm Chemistry

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.563[a]	.317	.306	14.2035

a. Predictors: (Constant), NSC_PhysicsPerc

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	p-value
1	Regression	6076.666	1	6076.666	30.121	.000[b]
	Residual	13113.085	65	201.740		
	Total	19189.751	66			

a. Dependent Variable: MdT_CHE103A

b. Predictors: (Constant), NSC_PhysicsPerc

Standard Error of Estimate (s_{YX})

- Measures the reliability of the estimating equation
- A measure of dispersion
- Measures the variability, or scatter of the observed values around the regression line

Interpreting the Standard Error of the Estimate

Measures the dispersion of the points around the regression line

If $s_e = 0$, equation is a "perfect" estimator

Is used to compute confidence intervals of the estimated value

Assumptions:

1) Observed *Y* values are normally distributed around each estimated value of *Y*

2) Constant variance

Assumptions of Linear Regression and Correlation

- Four major assumptions of regression
- Normality
- Homoscedasticity – Equality of variance for each x-value
- Independence of errors
- Linearity

Evaluating the Assumptions

- Homoscedasticity
  - Plot Studentized residuals of errors vs. independent variable
- Normality
  - Tally the Studentized residuals into a frequency distribution and create a histogram of results
- Independence
  - Plot residuals versus time in which they were collected
  - Also measured by Durbin-Watson statistic

Regression and Correlation Analyses: Limitations, Errors and Caveats

- Misuses:
  - Estimating beyond range of data (extrapolation)
  - Use to determine cause and effect relationship
  - Failure to recognize some variables can be dependent on time
  - Misrepresenting r^2 and r values
  - Finding relationships when they do not exist
- Caveats:
  - Know limitations of technique
  - Use common sense

Conclusion

1. Used scatter diagram to visualize relationship between two variables
2. Described the linear regression model
3. Explained ordinary least-squares method in generating equation
4. Used the estimated equation to estimate values
5. Learned the use of correlation analysis
6. Learned the limitations of regression and correlation analysis

References

Burns, R. B. & Burns, R. A. 2008. *Business research methods and statistics using SPSS*, London: SAGE Publications Ltd.

Levine D, Krehbiel T. C., and Berenson M. L., 2003. *Business Statistics: A First Course,* New York*:* Prentice Hall

Remenyi, D., Onofrei, G., English, J. 2009. *An introduction to Statistics using Microsoft Excel.* Reading: Academic Publishing Ltd.